DOCUMENT RESUME

ED 419 003 TM 028 300

AUTHOR van der Linden, Wim J.; Reese, Lynda M.

TITLE A Model for Optimal Constrained Adaptive Testing.

INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.

REPORT NO RR-97-01
PUB DATE 1997-00-00
NOTE 23p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; *Computer Assisted Testing; Computer Simulation; *Estimation (Mathematics); Foreign Countries; Higher Education; Selection; Test Construction; *Test Content; Test Items
IDENTIFIERS *Constraints; Law School Admission Test

ABSTRACT

A model for constrained computerized adaptive testing is proposed in which the information in the test at the ability estimate is maximized subject to a large variety of possible constraints on the contents of the test. At each item-selection step, a full test is first assembled to have maximum information at the current ability estimate fixing the items previously administered. Then the item with maximum information is selected from the test. All test assembly is optimal due to the use of a linear programming model that is automatically updated to allow for the attributes of items already administered as well as the new value of the ability estimator. A simulation study using a pool of 753 items from the Law School Admission Test (LSAT) showed that for adaptive tests of realistic lengths the ability estimator did not suffer any loss of efficiency from the presence of 433 constraints on the item selection process. (Contains 3 tables, 2 figures, and 35 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 419 003

TM

A Model for Optimal Constrained Adaptive Testing

Research Report 97-01

Wim J. van der Linden, University of Twente
and
Lynda M. Reese, Law School Admission Council

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028300

BEST COPY AVAILABLE

A Model for Optimal Constrained Adaptive Testing

Wim J. van der Linden
University of Twente

Lynda M. Reese
Law School Admission Council

Abstract

A model for constrained computerized adaptive testing is proposed in which the information in the test at the ability estimate is maximized subject to a large variety of possible constraints on the contents of the test. At each item-selection step, a full test is first assembled to have maximum information at the current ability estimate fixing the items previously administered. Then the item with maximum information is selected from the test. All test assembly is optimal due to the use of a linear programming model which is automatically updated to allow for the attributes of the items already administered as well as the new value of the ability estimator. A simulation study using a pool of 753 items from the LSAT showed that for adaptive tests of realistic lengths the ability estimator did not suffer any loss of efficiency from the presence of 433 constraints on the item selection process.

A Model for Optimal Constrained Adaptive Testing

The concept of adapting the difficulty of the test to the ability of the individual examinee is as old as the first intelligence test (Binet & Simon, 1905). In the Binet-Simon test, the items varied according to age group and the examiner was instructed to infer the next age group from the responses of the examinee to the previous test items until the true age group could be identified with sufficient certainty. In doing so, Binet and Simon intuitively followed the statistical principle that the information provided by test items is maximal if their difficulty matches the level of ability of the examinee.

Since modern group-based testing was introduced, attempts have been made to implement this principle of adaptivity in a practical format. One of the first attempts was two-stage testing--a testing format in which the score on a routing test directs the examinee to one of a limited number of measurement tests. In the self-scoring flexilevel test, a testing format proposed by Lord (1980, chap. 8), the examinee scores his/her own responses by scratching an answer sheet and is instructed to move on to the next item as a function of the correctness of the response. In Weiss' (1973) computerized stradaptive test, the items in the pool are divided into strata of difficulty and ordered according to their discrimination power within each stratum. The examinee moves to the next item in the higher stratum if his/her response is correct but to a lower stratum if it is incorrect. For a more extensive description of these early forms of adaptive testing, see Wainer (1990) or Weiss (1985).

With the advent of powerful personal computers and the acceptance of item response theory (IRT) as a tool for calibrating item pools, large-scale application of fully computerized adaptive testing (CAT) has become possible. A well-known procedure in adaptive testing is maximum-information item selection in combination with maximum-likelihood estimation of ability. In this paper, it is assumed that the responses to the items in the pool fit the three-parameter logistic (3-PL) response model

$$P_i(\theta) \equiv \text{Prob}\{U_i = 1\} \equiv c_i + (1-c_i)[1+\exp(-a_i(\theta_a - b_i))]^{-1}, \quad (1)$$

where $\theta_a \in (-\infty, \infty)$ is a parameter for the ability of examinee a, and $b_i \in (-\infty, \infty)$ and $a_i \in [0, \infty)$ are parameters for the difficulty and discrimination power of item i, respectively. For this model Fisher's information on θ in item i can be shown to be equal to

$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta), \quad (2)$$

with $Q_i(\theta) \equiv 1 - P_i(\theta)$. The maximum-information principle selects the next item to have a

maximum value for (2) at the current ability estimate. With a modern PC the time needed to calculate the maximum-likelihood ability estimate and select the item with maximum information from an item pool of realistic size is hardly noticeable by the examinee.

Paradoxically, now that fully computerized CAT is technically possible the interest seems to be moving back to earlier forms of adaptive testing. The reason for this unexpected development lies in the fact that the original conception of CAT focusses entirely on the statistical aspects of item selection and ability estimation and ignores all other test specifications typically in use in testing programs. As a consequence, it may lead to testing programs that:

1. do not guarantee equal composition of tests across examinees, and hence loose their face validity;
2. excludes the use of item pools with dependencies between the items, for example, between items that can not be administered in the same test because one item contains a clue to the solution to another item or between items that have to be presented in sets because they are linked to a common stimulus;
3. overexposes some items, with the potential danger that the items become known prematurely to the examinees;
4. do not allow for the possibility of reviewing responses to earlier items--a feature some programs want to offer to their examinees.

Several solutions to these problems have been proposed. Wainer and Kiely (1987) suggest adaptive testing from a pool of testlets rather than individual items, designing the testlets to ensure adequate content coverage in the individual tests. The same goal is addressed in the proposal by Kingsbury and Zara (1991) who suggest spiraling item selection along subsets of items in the pool defining relevant content dimensions. Adema (1990) and Luecht (1995) use optimization techniques to assemble a system of two-stage tests with each possible route meeting the same set of test specifications. Reese and Schnipke (1996) combine the ideas of two-stage and testlet-based testing. A probabilistic mechanism to govern the exposure rates of items in CAT is presented in Sympson and Hetter (1985). Stocking and Swanson (1993) propose a heuristic for sequential item selection that treats the test specifications as well as the goal of maximum information as "desirable properties" of the test and then compromises between them at each item-selection step.

It is the purpose of the present paper to propose a new form of constrained CAT. The procedure starts with the on-line assembly of a full test that meets all of the specifications and has maximum information at an initial estimate of the ability of the examinee. The assembly of the test is optimal due to the use of a linear programming (LP) model of the test

BEST COPY AVAILABLE

specifications. The first item to be administered is selected from this test according to the maximum-information principle. At each next step, the LP model is updated to allow for the values of the attributes of the items already administered, and the remaining part of the test is reassembled to have maximum information at the new ability estimate. The approach improves on conventional multi-stage or testlet-based adaptive testing designs in that there is no need to assemble fixed subtests or testlets in advance. All test assembly is on line to ensure maximum information at the current ability estimate. At the same time, unlike conventional CAT, item selection automatically satisfies the test specifications. The idea to base CAT on a process of reassembling full tests was developed independently by Cordova (1996). The approach is an alternative to the sequential heuristic proposed by Stocking and Swanson (1993); it is more rigorously based on the ideas developed for the application of linear programming (LP) to optimal test assembly, does guarantee that all of the test specifications are met, and has the explicit objective of maximum information in the test. A discussion of the precise differences between existing approaches and the present approach to constrained adaptive testing is postponed until the latter has been presented in more detail.

In the remaining part of the paper, constrained adaptive test assembly is first conceptualized as an adaptive solution to an LP model for test assembly. An example of a model is given and possible implementations are discussed. For two different implementations, the statistical properties of the ability estimator are compared in a simulation study using an existing item pool for the Law School Admission Test (LSAT).

General Model of Constrained Test Assembly

The concept underlying the following sections is that the process of test assembly can be characterized as an instance of constrained optimization. Formally, each constrained optimization problem has: (1) an objective function defined on the decision variables of the problem which is maximized or minimized; and (2) a series of constraints on the possible values of the decision variables which together define a feasible solution to the problem. In test assembly, for example, the objective may be to match the test information function to a target and the constraints may require that prespecified numbers of items be selected from certain content categories. If the objective function and constraints are linear in the decision variables, the problem belongs to the domain of linear programming (LP), which has a large body of algorithms and heuristics to solve its problems. A large variety of conventional test assembly problems have been shown to lend themselves to modeling as an LP problem with 0-1 decision variables. Some relevant references are: Adema (1992a, 1992b), Adema, Boekkooi-Timminga and van der Linden (1991), Adema and van der Linden (1989), Amstrong and Jones (1992), Amstrong, Jones and Wu (1992), Boekkooi-Timminga (1987,

1990), Theunissen (1985, 1986), Timminga and Adema (1995, 1996), van der Linden (1994; to appear), van der Linden and Boekkooi-Timminga (1988), and van der Linden and Luecht (1996).

An important distinction in test assembly is the one between constraints on categorical and quantitative attributes of test items. Categorical attributes introduce a partitioning of the item pool with different subsets of items corresponding to different levels of the attribute. Some examples of categorical attributes are: item content, cognitive level, item format, and gender orientation. A quantitative attribute is a parameter or coefficient with possibly different numerical values for each item. Examples of this type of attribute are: item p-value, expected response time, and item exposure rate. Constraints may also be needed to guarantee that items linked to the same stimulus are administered as sets. In addition, these stimuli themselves may involve constraints on categorical (e.g., content classification) or quantitative attributes (e.g., word count).

The problem of constrained CAT can now be represented as a series of updates of the following optimization problem:

maximize information at current ability estimate **(2)**

subject to possible constraint(s) on the

length of the test; **(3)**
number of item sets in the test; **(4)**
number(s) of items per item set; **(5)**
categorical item attributes; **(6)**
quantitative item attributes; **(7)**
dependencies between items in sets; **(9)**
categorical item set attributes; **(10)**
quantitative item set attributes. **(11)**

In addition, a few technical constraints may be necessary to solve the optimization problem. The following section gives an example of an LP formulation of this verbally stated problem.

Example

To present the example, the following definitions are needed: The items in the pool are indexed by i=1,...,I. In addition, the pool is assumed to consists of item sets, V_j, j=1,...,J, each of which may have a different number of items. For each item a decision variable x_i is used which takes the value 1 if the item is included in the test and the value 0 otherwise.

Likewise, a second decision variable z_j is used to decide whether (z_j=1) or not (z_j=0) item set j is included in the test. In addition, the exemplary attributes in Table 1 are used.

Table 1

Exemplary Item and Item Set Attributes

Attribute	Value
Cognitive Level of Item	Reading Comprehension (C_1); Analytic Reasoning (C_2); Logical Reasoning (C_3)
Expected Response Time for Item	$r_i \in (0, \infty)$
Frequency of Previous Item Usage	$f_i \in \{0,1,\ldots\}$
Content of Item Set	Humanities (S_1); Social Sciences (S_2)

The following example of the test assembly problem is given:

$$\text{maximize} \sum_{i=1}^{I} I(\hat{\theta}) x_i \qquad \text{(maximum information at } \hat{\theta}\text{)} \qquad \textbf{(12)}$$

subject to

$$\sum_{i=1}^{I} x_i = n, \qquad \text{(test length)} \qquad \textbf{(13)}$$

$$\sum_{j=1}^{J} z_j = m, \qquad \text{(number of item sets)} \qquad \textbf{(14)}$$

$$\sum_{i \in V_j} x_i \le n_j^{(u)} z_j, \quad j=1,\ldots,J, \qquad \text{(number of items in item set j)} \qquad \textbf{(15)}$$

$$\sum_{i \in V_j} x_i \ge n_j^{(l)} z_j, \quad j=1,\ldots,J, \qquad \text{(number of items in item set j)} \qquad \textbf{(16)}$$

$$\sum_{i \in C_h} x_i \le n_h^{(u)}, \quad h=1, 2, 3, \qquad \text{(number of items per cognitive level)} \qquad \textbf{(17)}$$

$$\sum_{i \in C_h} x_i \ge n_h^{(l)}, \quad h=1, 2, 3, \qquad \text{(number of items per cognitive level)} \qquad \textbf{(18)}$$

$$\sum_{i=1}^{I} r_i x_i \leq r^{(u)} \qquad \text{(response time available)} \qquad \textbf{(19)}$$

$$f_i x_i \leq f^{(u)}, \quad i=1,\ldots,I, \qquad \text{(maximum item exposure)} \qquad \textbf{(20)}$$

$$\sum_{j \in S_g} z_j \leq n_g^{(u)}, \quad g=1, 2 \qquad \text{(number of item sets per content category)} \qquad \textbf{(21)}$$

$$\sum_{j \in S_g} z_j \geq n_g^{(l)}, \quad g=1, 2 \qquad \text{(number of item sets per content category)} \qquad \textbf{(22)}$$

$$x_{31} + x_{32} + x_{33} + x_{34} \leq 1 \qquad \text{(mutually exclusive items)} \qquad \textbf{(23)}$$

$$z_8 + z_9 + z_{10} \leq 1 \qquad \text{(mutually exclusive item sets)} \qquad \textbf{(24)}$$

$$x_i = 0, 1, \quad i=1,\ldots,I, \qquad \text{(domain of decision variables)} \qquad \textbf{(25)}$$

$$z_j = 0, 1, \quad j=1,\ldots,J. \qquad \text{(domain of decision variables)} \qquad \textbf{(26)}$$

The right-hand side coefficients in the constraints are bounds on numbers of items (n) or item sets (m). Upper and lower bounds are denoted by a corresponding superscript. Note that some of the constraints are formulated using the decision variables for the items (x_i) and others using the variables for the item sets (z_j). The constraints in (15)-(16) have both types of variables to ensure that individual items in sets are chosen if and only if a sufficient number from their sets are chosen. It is evident that the model only has a solution if the numbers in the right-hand side coefficients are chosen consistently and the pool has enough items to satisfy these numbers. These conditions are assumed to be met in a deliberately designed CAT program.

The model in (12)-(26) is equivalent to the maximin model for test assembly (van der Linden and Boekkooi-Timminga, 1989), with the exception that it does not maximize the information in the test proportionally at a number of θ values but at an estimate of θ for a single examinee. A review of the constraints available to model a large variety of test specifications is given in the same paper.

Models for test assembly as in (12)-(26) can be solved for an optimal test (=set of values for the decision variables) using a standard software package for LP or a choice from the algorithms and heuristics offered in the test assembly package ConTEST (Timminga, van der Linden & Schweizer, 1996). For test assembly models with the special structure of a

BEST COPY AVAILABLE

network-flow problem, efficient algorithms are possible (Amstrong, Jones & Wu, 1992). Typically, the use of each of these algorithms is preceded by some form of preprocessing of the model or the item pool; for example, a solution of a model with a constraint as in (19) is generally obtained quicker if all items with $f_i > f^{(u)}$ are first removed from the pool.

The next section discusses how to implement models as in (12)-(26) in a CAT program.

Adaptive Implementation of the Model

It is assumed that the test stops as soon as n items are administered. Other stopping rules are possible but this rule is believed to enhance the face validity of the test. Adaptive implementation of the model in (12)-(26) involves the on-line execution of the following steps for each examinee:

Step 1: Initialize the model;
Step 2: Assemble an initial test according to the model;
Step 3: Administer the item with maximum information at the ability estimate;
Step 4: Update the model;
Step 5: Reassemble the remaining part of the test putting the items not administered back into the pool;
Step 6: Repeat Steps 3-6 until n items have been administered.

The algorithm is adaptive because of Step 4. The update of the model in this step involves both an update of $\hat{\theta}$ in the objective function in (12) and an update to allow for the attributes of the item administered. The only thing needed to perform the latter is to insert a constraint into the model that sets the decision variable of this item equal to 1. For example, if Item 22 is selected, the constraint $x_{22}=1$ is inserted.

Note that when reassembling the remaining part of the test in Step 5, the items not yet administered are put back into the pool. Hence, the newly assembled part of the test is always at least as good as the old part but most likely better since the ability estimate has been updated. Also, if a feasible solution to the model exists for the initial test, the problem of reassembling later parts of the test remains feasible.

In Table 2 the algorithm is illustrated for a 5-item test. The items in the upper triangle are the items already administered. The items in the lower triangle form the part of the test reassembled using the updated model (Step 5). The bold numbers in this triangle are the items selected according to the maximum-information principle. Note that bold numbers are moved to the upper triangle in the next column of the table.

Table 2

Example of a 5-item constrained adaptive test

Selection of Item	#1	#2	#3	#4	#5
	--	39	39	39	39
	13	--	14	14	14
	27	8	--	41	41
	28	***14***	22	--	22
	39	41	37	***22***	--
	41	49	***41***	37	***6***

Note. Numbers in upper triangle are items already administered. Italic numbers in lower triangle are items in the reassembled part of the test. Bold numbers are items selected according to the maximum-information principle.

Possible initializations of the model.

How the model should be initialized in Step 1 has not yet been explained. An obvious way to do so is to choose a plausible value for $\hat{\theta}$ based on knowledge of the ability distribution of the population of examinees and to choose the values for the bounds in the constraints on the basis of the test specifications. A more sophisticated initialization of $\hat{\theta}$ is to choose a value based on prior information on the values of relevant background variables for the examinee. A method for estimating θ directly from background variables is presented in van der Linden (submitted). An alternative is to choose a prior value for $\hat{\theta}$ and administer a short CAT as a pretest, ignoring the constraints in the model. The suggestions is based on the observation that the presence of large numbers of constraints in the test assembly models may slow down the convergence of the ability estimator. Therefore, it may be advantageous to relax the algorithm first and impose the constraints on the item selection process when the ability estimator has had some time to stabilize. Stabilization has been shown to be remarkably quick for a Bayesian alternative to the maximum-information principle of item selection known as the Maximum Predicted Posterior Expected Information Criterion (van der Linden, 1996). If the constraints are introduced at a later moment in the test, the decision variables of the items already administered have to be fixed at 1. Of course, to keep the original model feasible, the pretest can not be longer than the smallest upper bound in the right-hand sides of the constraints on item numbers in the model.

BEST COPY AVAILABLE

Item Sets and Item Review

The presence of items sets in the pool entails no special measures as long as the structure of the pool has been modeled correctly by constraints such as those in (14)-(15), (21)-(22), and (24) in the exemplary model. If an item set is chosen, an optimal number of items in the set between the given bounds is also chosen. Normally, item sets are to be administered intact. If so, Step 4 and 5 in the algorithm are postponed until the last item in the set has been administered. In the (unlikely) case that the items need not be administered as an intact set, the procedure can just be continued and the algorithm automatically selects the right number of items from the set at optimal moments.

If the examinees are given the opportunity to review their responses within blocks of items, the only possible consequence is a revision of the ability estimate if some of the responses are changed. Thus, when moving to a next block, $\hat{\theta}$ may have to be revised but the set of constraints in the model need not be updated.

Statistical Properties of the Ability Estimator

To study the effect of constraints in the adaptive item selection process on the ability estimator for a realistic adaptive testing program, a simulation study was run using a pool of 753 items from the LSAT. The pool consisted of three different sections, which are labeled here as SA, SB, and IA. All items were calibrated using the 3-PL model given in (1). The length of the adaptive test was set equal to n=50, with the following distribution of items across sections: SA: 12 items; SB: 14 items; and IA: 24 items. Large numbers of linear constraints were imposed on the item selection process to deal with the item-set structure of the pool as well as existing specifications with respect to item (sub)types, types of stimuli in item sets, gender and minority orientation of the stimuli, answer key distributions, and words counts. The numbers of decision variables and constraints in the model for the complete test as well as its three sections are given in Table 3.

Table 3

Numbers of items, item sets, decision variables, and constraints in the model

Level	#Items	#Item Sets	#Variables	#Constraints
Test	753	3	804	433
SA	208	24	232	179
SB	240	24	264	218
IA	305	0	305	30

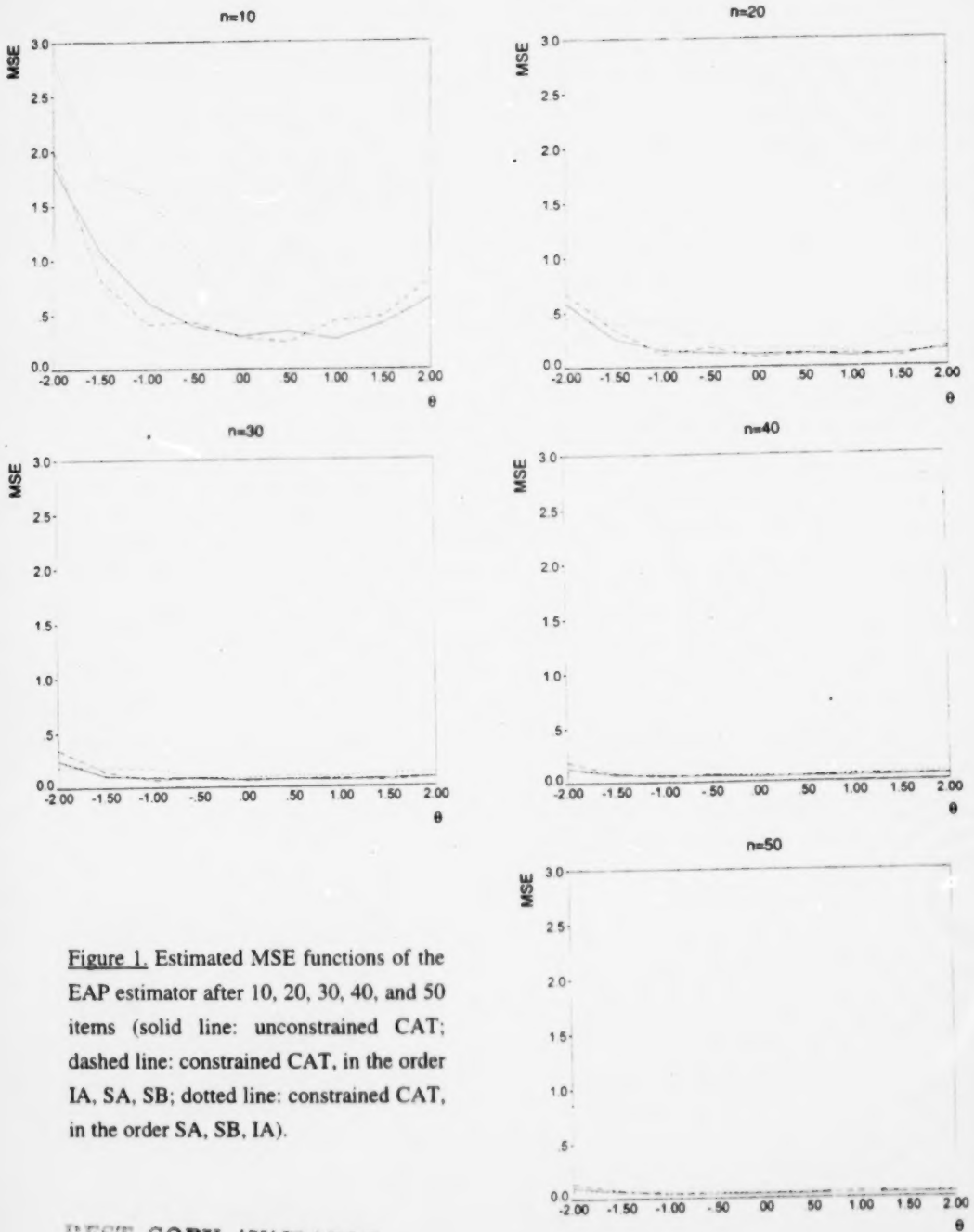

Figure 1. Estimated MSE functions of the EAP estimator after 10, 20, 30, 40, and 50 items (solid line: unconstrained CAT; dashed line: constrained CAT, in the order IA, SA, SB; dotted line: constrained CAT, in the order SA, SB, IA).

BEST COPY AVAILABLE

The following three different conditions were simulated:

1. Constrained CAT, with sections in the order IA, SA, SB;
2. Constrained CAT, with sections in the order SA, SB, IA;
3. Unconstrained CAT.

Because Section IA was least severely constrained, a comparison between the results for the first two conditions shows the effect of imposing the majority of the constraints after the ability estimator is stabilized. The comparison between the first two and the last condition shows the effect of the 433 constraints on the ability estimator.

Adaptive tests were simulated for θ=-2.0,-1.5,...,2.0, and the procedure was replicated 100 times for each θ value. Ability was estimated using the EAP estimator with a uniform prior distribution. The initial ability estimate was set equal to 0. At each step the LP model was solved using the First Acceptable Integer Solution Algorithm (Adema, 1992b; Timminga, van der Linden & Schweizer, 1996, sect. 6.6). This heuristic is based on the following adaptation of the branch-and-bound method. Let z_{LP} be the value of the objective function in the solution to the relaxed model. This value is as an upper bound to the solution of the model with 0-1 variables. The branch-and-bound search is stopped as soon as the current solution is larger than $h_1 z_{LP}$, with $h_1<1$ but large enough to guarantee a satisactory result. In addition, following Crowder, Johnson, and Padberg (1983), the optimal reduced costs in the relaxed solution are used to fix some of the nonbasic variables. Let d_j be the costs associated with nonbasic variable x_j. Then, if $x_j=0$ in the relaxed solution and $z_{LP}-h_2 z_{LP}<d_j$, $h_2<1$, the variable is fixed to 0. Likewise, x_j is fixed to 1 if $x_j=1$ in this solution and $z_{LP}-h_2 z_{LP}<-d_j$. For the LP models in the present example, the best setting found was h_1=.90 and h_2=.91. Parameter h_2 has to be set larger than h_1, but if it is set too high, overconstraining may occur. In manual test assembly, the heuristic is then rerun with a lower value for this parameter. In the current framework of adaptive testing, however, it was decided not to reassemble the test and to select the next item simply from the last test assembled. The effect of this measure, which was applied for 4.06% of all items selected in this study, is possibly less than optimal item selection and hence underestimation of the efficiency of the ability estimator. The results from the comparison between the mean-squared error (MSE) of the ability estimator in the constrained and unconstrained adaptive modes presented below is therefore expected to be slightly conservative with respect to the former.

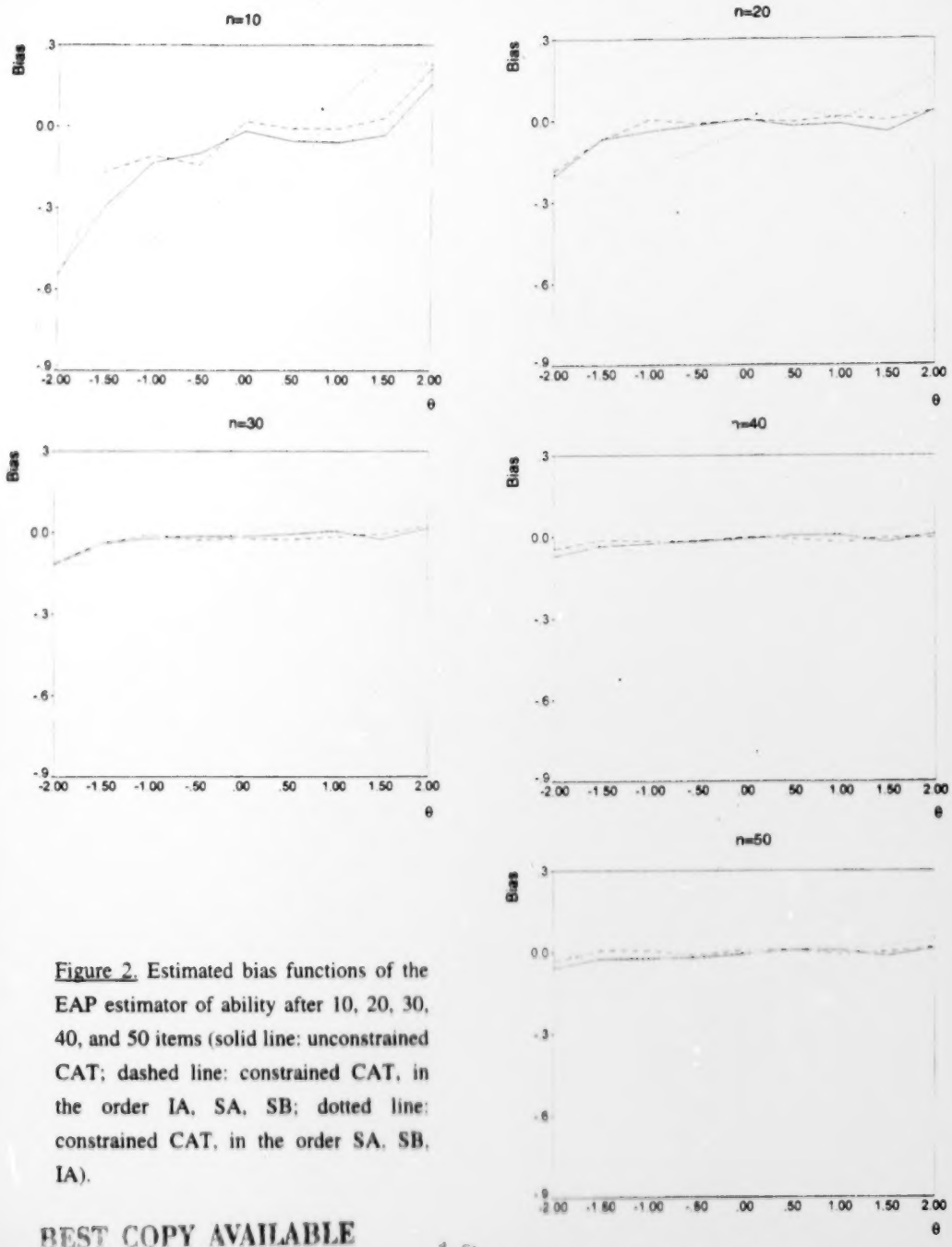

Figure 2. Estimated bias functions of the EAP estimator of ability after 10, 20, 30, 40, and 50 items (solid line: unconstrained CAT; dashed line: constrained CAT, in the order IA, SA, SB; dotted line: constrained CAT, in the order SA, SB, IA).

BEST COPY AVAILABLE

All runs were made on a PC with Pentium/133MHz processor. The CPU times needed to select an item in the constrained mode, that is, to update $\hat{\theta}$, reassemble the test, and select an item with maximum information from it, were all within 1-2 secs. These figures show that the approach proposed in this paper is practically feasible for item pools and test specifications such as those used in this example.

The MSE functions of the EAP ability estimator after n=10, 20, 30, 40, and 50 are presented in Figures 1. For n=10, the functions for Condition 1 (constrained CAT, with order: IA, SA, SB) and Condition 3 (unconstrained CAT) show about equal results for all values of θ. The function for Condition 2 reveals relatively poor performance for the CAT version with a more severely constrained section at the beginning of the test. However, the effect is already small when 20 items are administered, and for more than 30 items the results for the three conditions are identical for all practical purposes. The bias functions in Figure 2 show the same pattern. Note that in both figures the results for the lower end of the θ scale tend to be somewhat poorer than those for the upper end. This difference in performance is likely to be due to underrepresentation of some categories of items at the lower end of the scale in the item pool.

Discussion

As already observed, other adaptive testing formats that can be used to deal with constraints on test contents are multi-stage and testlet-based adaptive testing. In multi-stage testing, the content of the test is adapted only at the end of previously determined stages. In addition, at each stage only a limited number of options is available each designed to be optimal for a previously selected ability level. In contrast, the present format adapts the content of the test to the updated ability estimate after each new item, selects the remaining part of the test from all options feasible for the item pool, and guarantees maximum information. Testlet-based adaptive testing offers more flexibility than multi-stage testing but in principle the same differences hold. In the Stocking and Swanson (1993) approach, all test specifications and the objective of maximal information are combined into a weighted objective function. Next, the items are selected from the pool to optimize this function in a sequential mode. Applying the approach to the empirical example in this paper, weights would have to be specified to reflect the desirability of each of the 433 constraints in the model. As a consequence of this complexity, unpredictable violations of the constraints as well as the principle of maximum information may occur. The approach in this paper, however, requires all constraints to be met. In addition, it is not based on sequential selection of single items but at each step selects all remaining items simultaneously to have maximum information at the ability estimate.

References

Adema, J.J. (1990). The construction of customized two-staged tests. Journal of Educational Measurement, 27, 241-253.

Adema, J.J. (1992a). Methods and models for the construction of weakly parallel tests. Applied Psychological Measurement, 16, 53-63.

Adema, J.J. (1992b). Implementations of the branch-and-bound method for test construction. Methodika, 6, 99-117.

Adema, J.J., Boekkooi-Timminga, E., & van der Linden, W.J. (1991). Achievement test construction using 0-1 linear programming. European Journal of Operations Research, 55, 103-111.

Adema, J.J. & van der Linden, W.J. (1989). Algorithms for computerized test construction using classical item parameters. Journal of Educational Statistics, 14, 279-290.

Amstrong, R.D. and Jones, D.H. (1992). Polynomial algorithms for item matching. Applied Psychological Measurement, 16, 365-373.

Amstrong, R.D., Jones, D.H., & Wu, I.-L. (1992). An automated test development of parallel tests. Psychometrika, 57, 271-288.

Binet, A., & Simon, Th.A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. l'Anneé Psychologie, 11, 191-336.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. Methodika, 1, 1101-112.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. Journal of Educational Statistics, 15, 129-145.

Cordova, M.J. (1996). Optimizations methods in computerized adaptive testing. Unpublished doctoral dissertation proposal, Rutgers University, New Brunswick, NJ.

Crowder, H., Johnson, E.L., & Padberg, H. (1983). Solving large-scale zero-one programming problems. Operations Research, 31, 803-834.

Kingsbury, G.G. & Zara, A.R. (1991). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2, 359-375.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Luecht, R.M., & Nungester, R.J. (1996). Some practical examples of computerized adaptive sequential testing (Internal Report). Philadelphia, PA: National Board of Medical Examiners.

Reese, L.M., & Schnipke, D.L. (1996, June). An evaluation of a two-stage testlet design for computerized adaptive testing. Paper presented at the annual meeting of the Psychometric Society, Banff, Alberta, Canada.

Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17, 277-292.

Sympson, J.M., & Hetter, R.D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, T.J.J.M. (1986). Optimization algorithms in test design. Applied Psychological Measurement, 10, 381-389.

Timminga, E. & Adema, J.J. (1995). Test construction from item banks (pp. 111-127). In G. H. Fischer & I.W. Molenaar (Eds.), The Rasch model: Foundations, recent developments, and applications. New York: Springer-Verlag.

Timminga, E. & Adema, J.J. (1996). An interactive approach to modifying infeasible 0-1 linear programming models for test construction. In G. Engelhard & M. Wilson (Ed.), Objective measurement: Theory into practice (Vol.3, pp. 419-436). Norwood, New Jersey: Ablex Publishing Company.

Timminga, E., van der Linden, W.J., & Schweizer, D.A. (1996). ConTEST [Computer program and manual]. Groningen, The Netherlands: iec ProGAMMA.

van der Linden, W.J. (1994). Optimum design in item response theory: Applications to test assembly and item calibration. In G.H. Fischer & D. Laming (Eds,), Contributions to mathematical psychology, psychometrics, and methodology (pp. 308-318). New York: Springer-Verlag.

van der Linden, W.J. (Ed.) (to appear). Optimal test assembly. Applied Psychological Measurement [Special issue].

van der Linden, W.J., & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subsets method. Applied Psychological Measurement, 12, 201-209.

van der Linden, W.J. (1996). Bayesian item selection criteria for adaptive testing (Research Report 96-1). Enschede, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.

van der Linden, W.J. (submitted). A procedure for empirical initialization of adaptive testing algorithms.

van der Linden, W.J., & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subsets method. Applied Psychological Measurement, 12, 201-209.

van der Linden, W.J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 17, 237-247.

van der Linden, W.J., & Luecht, R.M. (1996). An optimization model for test assembly to match observed-score distributions. In G. Engelhard & M. Wilson (Ed.), Objective measurement: Theory into practice (Vol.3, pp. 405-418). Norwood, New Jersey: Ablex Publishing Company.

Wainer, H. (Ed.) (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: 1990.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Weiss, D.J. (1973). The stradaptive computerized test ability test (Research Report No. 73-3). Minneapolis: University of Twente, Department of Psychology.

Weiss, D.J. (1985). Adaptive testing by computer. Journal of Counseling and Clinical Psychology, 53, 774-789.

Authors' Note

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the authors and do not necessarily reflect the position or policy of LSAC. This report is also published in the LSAC Research Report Series. The authors are indebted to Wim M.M. Tielen for writing the simulation program and to David A. Schweizer for adapting the CONSOL software.

Address all correspondence to: W.J. van der Linden, Faculty of Educational Science and Technology, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: vanderlinden@edte.utwente.nl.

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-97-01	W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*
RR-96-04	C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*
RR-96-03	C.A.W. Glas, *Testing the Generalized Partial Credit Model*
RR-96-02	C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*
RR-96-01	W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*
RR-95-03	W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*
RR-95-02	W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*
RR-95-01	W.J. van der Linden, *Some decision theory for course placement*
RR-94-17	H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*
RR-94-16	H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*
RR-94-15	H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*
RR-94-13	W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*
RR-94-12	R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*
RR-94-10	W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*
RR-94-9	L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*
RR-94-8	R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*
RR-94-7	W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*
RR-94-6	W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*
RR-94-5	R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*
RR-94-4	M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.

BEST COPY AVAILABLE